Topic: Multivariate Analysis

Course: ***M.A. Geography (Sem.-3)***

Paper: M-301(CC-10)

Quantitative Techniques and Research Methodology

By

Dr. Md. Nazim

Professor, Department of Geography

Patna College, Patna University

Contact No- 8409110509

Email – dr.nazim2011@gmail.com

Lecture- 9

Concept:

***Multivariate analysis** (**MVA**) is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. Typically, MVA is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important. A modern, overlapping categorization of MVA includes:*

- *Normal and general multivariate models and distribution theory*
- *The study and measurement of relationships*
- *Probability computations of multidimensional regions*
- *The exploration of data structures and patterns*

Multivariate analysis can be complicated by the desire to include physics-based analysis to calculate the effects of variables for a hierarchical "system-of-systems". Often, studies that wish to use multivariate analysis are stalled by the dimensionality of the problem. These concerns are often eased through the use of surrogate models, highly accurate approximations of the physics-based code. Since surrogate models take the form of an equation, they can be evaluated very quickly. This becomes an enabler for large-scale MVA studies: while a Monte Carlo simulation across the design space is difficult with physics-based codes, it becomes trivial when evaluating surrogate models, which often take the form of response-surface equations.

Multivariate analysis methods are used in the evaluation and collection of statistical data to clarify and explain relationships between different variables that are associated with this data.

Multivariate tests are always used when more than three variables are involved and the context of their content is unclear. The goal is to both detect a structure, and to check the data for structures.

History:

Anderson's 1958 textbook, An Introduction to Multivariate Statistical Analysis, educated a generation of theorists and applied statisticians; Anderson's book emphasizes hypothesis testing via likelihood ratio tests and the properties of power functions: Admissibility, unbiasedness and monotonicity.MVA once solely stood in the statistical theory realms due to the size, complexity of underlying data set and high computational consumption. With the dramatic growth of computational power, MVA now plays an increasingly important role in data analysis and has wide application in OMICS fields.

Types of multivariate analysis methods:

Multivariate methods can be subdivided according to different aspects. First of all, they are differentiated according to whether the aim is to discover a structure within the combination of data, or whether the data is to be checked with a certain structure. a structure The structure-determining methods include:

Factor analysis: *Reduces the structure to relevant data and individual variables. Factor studies focus on different variables, so they are further subdivided into main component analysis and correspondence analysis. For example: Which website elements have the greatest influence on purchasing behavior?*

Cluster analysis: *Observations are graphically assigned to individual variable groups and classified on the basis of these. The results are clusters and segments, such as the number of buyers of a particular product, who are between 35 and 47 years old and have a high income.*

Regression Analysis: *Investigates the influence of two types of variables on each other. Dependent and nondependent variables are spoken of. The former are so-called explanatory variables, while the latter are explanatory variables. The first describes the actual state on the basis of data, the second explains this data by means of dependency relationships between the two variables. In practice, several changes of web page elements correspond to independent variables, while the effects on the conversion rate would be the dependent variable.*

Variance analysis: *Determines the influence of several or individual variables on groups by calculating statistical averages. Here you can compare variables within a group as well as different groups, depending on where deviations are to be assumed. For example: Which groups most often click on the' Buy Now' button in your shopping cart?*

Discriminant analysis: *Used in the context of variance analysis to differentiate between groups that can be described by similar or identical characteristics. For example, by which variables do different groups of buyers differ?*

In order to understand multivariate analysis, it is important to understand some of the terminology. A variate is a weighted combination of variables. The purpose of the analysis is to find the best combination of weights. Nonmetric data refers to data that are either qualitative or categorical in nature. Metric data refers to data that are quantitative, and interval or ratio in nature.

Initial Step—Data Quality:

Before launching into an analysis technique, it is important to have a clear understanding of the form and quality of the data. The form of the data refers to whether the data are nonmetric or metric. The quality of the data refers to how normally distributed the data are. The first few techniques discussed are sensitive to the linearity, normality, and equal variance assumptions of the data. Examinations of distribution, skewness, and kurtosis are helpful in examining distribution. Also, it is important to understand the magnitude of missing values in observations and to determine whether to ignore them or impute values to the missing observations. Another data quality measure is outliers, and it is important to determine whether the outliers should be removed. If they are kept, they may cause a distortion to the data; if they are eliminated, they may help with the assumptions of normality. The key is to attempt to understand what the outliers represent.

Multiple Regression Analysis:

Multiple regression is the most commonly utilized multivariate technique. It examines the relationship between a single metric dependent variable and two or more metric independent variables. The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity, and equal variance are carefully observed. The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly. Multiple regression is often used as a forecasting tool.

Logistic Regression Analysis:

Sometimes referred to as "choice models," this technique is a variation of multiple regression that allows for the prediction of an event. It is allowable to utilize nonmetric (typically binary) dependent variables, as the objective is to arrive at a probabilistic assessment of a binary choice. The independent variables can be either discrete or continuous. A contingency table is produced, which shows the classification of observations as to whether the observed and predicted events match. The sum of events that were predicted to occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events, is a measure of the effectiveness of the model. This tool helps predict the choices consumers might make when presented with alternatives.

Discriminant Analysis:

The purpose of discriminant analysis is to correctly classify observations or people into homogeneous groups. The independent variables must be metric and must have a high degree of normality. Discriminant analysis builds a linear discriminant function, which can then be used to classify the observations. The overall fit is assessed by looking at the degree to which the group means differ (Wilkes Lambda or D2) and how well the model classifies. To determine which variables have the most impact on the

discriminant function, it is possible to look at partial F values. The higher the partial F, the more impact that variable has on the discriminant function. This tool helps categorize people, like buyers and nonbuyers.

Multivariate Analysis of Variance (MANOVA):

This technique examines the relationship between several categorical independent variables and two or more metric dependent variables. Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means and F tests between three or more means), MANOVA examines the dependence relationship between a set of dependent measures across a set of groups. Typically this analysis is used in experimental design, and usually a hypothesized relationship between dependent measures is used. This technique is slightly different in that the independent variables are categorical and the dependent variable is metric. Sample size is an issue, with 15-20 observations needed per cell. However, too many observations per cell (over 30) and the technique loses its practical significance. Cell sizes should be roughly equal, with the largest cell having less than 1.5 times the observations of the smallest cell. That is because, in this technique, normality of the dependent variables is important. The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined.

Factor Analysis:

When there are many variables in a research design, it is often helpful to reduce the variables to a smaller set of factors. This is an independence technique, in which there is no dependent variable. Rather, the researcher is looking for the underlying structure of the data matrix. Ideally, the independent variables are normal and continuous, with at least three to five variables loading onto a factor. The sample size should be over 50 observations, with over five observations per variable. Multicollinearity is generally preferred between the variables, as the correlations are key to data reduction. Kaiser's Measure of Statistical Adequacy (MSA) is a measure of the degree to which every variable can be predicted by all other variables. An overall MSA of .80 or higher is very good, with a measure of under .50 deemed poor.

There are two main factor analysis methods: common factor analysis, which extracts factors based on the variance shared by the factors, and principal component analysis, which extracts factors based on the total variance of the factors. Common factor analysis is used to look for the latent (underlying) factors, whereas principal component analysis is used to find the fewest number of variables that explain the most variance. The first factor extracted explains the most variance. Typically, factors are extracted as long as the eigenvalues are greater than 1.0 or the Scree test visually indicates how many factors to extract. The factor loadings are the correlations between the factor and the variables. Typically a factor loading of .4 or higher is required to attribute a specific variable to a factor. An orthogonal rotation assumes no correlation between the factors, whereas an oblique rotation is used when some relationship is believed to exist.

Cluster Analysis:

The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. Outliers are a problem with this technique, often caused by too many irrelevant variables. The sample should be representative of the population, and it is desirable to have uncorrelated factors. There are three main clustering methods: hierarchical, which is a treelike process appropriate for smaller data sets; nonhierarchical, which requires specification of the number of clusters a priori; and a combination of both. There are four main rules for developing clusters: the clusters should be different, they should be reachable, they should be measurable, and the clusters should be profitable (big enough to matter). This is a great tool for market segmentation.

Multidimensional Scaling (MDS):

The purpose of MDS is to transform consumer judgments of similarity into distances represented in multidimensional space. This is a decompositional approach that uses perceptual mapping to present the dimensions. As an exploratory technique, it is useful in examining unrecognized dimensions about products and in uncovering comparative evaluations of products when the basis for comparison is unknown. Typically there must be at least four times as many objects being evaluated as dimensions. It is possible to evaluate the objects with nonmetric preference rankings or metric similarities (paired comparison) ratings. Kruskal's Stress measure is a "badness of fit" measure; a stress percentage of 0 indicates a perfect fit, and over 20% is a poor fit. The dimensions can be interpreted either subjectively by letting the respondents identify the dimensions or objectively by the researcher.

Correspondence Analysis:

This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings. However, unlike MDS, both independent variables and dependent variables are examined at the same time. This technique is more similar in nature to factor analysis. It is a compositional technique, and is useful when there are many attributes and many companies. It is most often used in assessing the effectiveness of advertising campaigns. It is also used when the attributes are too similar for factor analysis to be meaningful. The main structural approach is the development of a contingency (crosstab) table. This means that the form of the variables should be nonmetric. The model can be assessed by examining the Chi-square value for the model. Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.

Conjoint Analysis:

Conjoint analysis is often referred to as "trade-off analysis," since it allows for the evaluation of objects and the various levels of the attributes to be examined. It is both a

compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed. A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level. Models can be built that identify the ideal levels and combinations of attributes for products and services.

Canonical Correlation:

The most flexible of the multivariate techniques, canonical correlation simultaneously correlates several independent variables and several dependent variables. This powerful technique utilizes metric independent variables, unlike MANOVA, such as sales, satisfaction levels, and usage levels. It can also utilize nonmetric categorical variables. This technique has the fewest restrictions of any of the multivariate techniques, so the results should be interpreted with caution due to the relaxed assumptions. Often, the dependent variables are related, and the independent variables are related, so finding a relationship is difficult without a technique like canonical correlation.

Structural Equation Modeling:

Unlike the other multivariate techniques discussed, structural equation modeling (SEM) examines multiple relationships between sets of variables simultaneously. This represents a family of techniques, including LISREL, latent variable analysis, and confirmatory factor analysis. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis. For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures. These tools are often used to evaluate many scaled attributes or to build summated scales.

Conclusions:

Each of the multivariate techniques described above has a specific type of research question for which it is best suited. Each technique also has certain strengths and weaknesses that should be clearly understood by the analyst before attempting to interpret the results of the technique. Current statistical packages (SAS, SPSS, S-Plus, and others) make it increasingly easy to run a procedure, but the results can be disastrously misinterpreted without adequate care.

Applications:

As a quantitative method, multivariate analysis is one of the most effective methods of testing usability. At the same time, it is very complex and sometimes cost-intensive. Software can be used to help, but the tests as such are considerably more complex than A/B tests in terms of study design. The decisive advantage lies in the number of

variables that can be considered and their weighting as a measure of the significance of certain variables.

Even four different versions of an article's headline can result in completely different click rates. The same applies to the design of buttons or the background color of the order form. In individual cases, it is therefore worth considering from a multivariate perspective also financially, especially for commercially oriented websites, such as online shops or websites, which are to be amortized through advertising.

- *Multivariate hypothesis testing*
- *Dimensionality reduction*
- *Latent structure discovery*
- *Clustering*
- *Multivariate regression analysis*
- *Classification and discrimination analysis*
- *Variable selection*
- *Multidimensional Scaling*
- *Data mining*

Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding. For instance, in analyzing financial instruments, the relationships among the various characteristics of the instrument are critical. In biopharmaceutical medicine, the patient's multiple responses to a drug need be related to the various measures of toxicity. Some of what falls into the rubric of multivariate analysis parallels traditional univariate analysis; for example, hypothesis tests that compare multiple populations. However, a much larger part of multivariate analysis is unique to it; for example, measuring the strength of relationships among various measurements.

Although there are many practical applications for each of the methods discussed in this overview, we cite some applications for the classification and discrimination methods in Sect. 6.5. The goal is to distinguish between two populations, or to classify a new observation in one of the populations. Examples are (a) solvent and insolvent companies based on several financial measures; (b) nonulcer dyspeptics versus normal individuals based on measures of anxiety, dependence, guilt, and perfectionism; (c) Alaskan vs. Canadian salmon based on measures of the diameters of rings. For other such applications, see Johnson and Wichern (1999).

Multivariate analysis, due to the size and complexity of the underlying data sets, requires much computational effort. With the continued and dramatic growth of computational power, multivariate methodology plays an increasingly important role in data analysis, and multivariate techniques, once solely in the realm of theory, are now finding value in applications. The data may be metrical, categorical, or a mixture of the two. Multivariate data may be, first, summarized by looking at the pair-wise associations. Beyond that, the different methods available are designed to explore and elucidate different features of the data. The article briefly summarizes the scope and

purpose of the following methods: cluster analysis, multidimensional scaling, principal components analysis, latent class analysis, latent profile analysis, latent trait analysis, factor analysis, regression analysis, discriminant analysis, path analysis, correspondence analysis, multilevel analysis, and structural equation analysis.

1. ***Breast Carcinoma Diagnosis***
2. ***Rapid Real-Time Raman Spectroscopy and Imaging-Guided Confocal Raman Spectroscopy for In Vivo Skin Evaluation and Diagnosis:***
3. ***Trisectionectomy:***
4. ***Retransplantation:***
5. ***Cancers of the Vulva and Vagina:***
6. ***Volatile Organic Compounds in Human Breath: Biogenic Origin and Point-of-Care Analysis Approaches:***

References:

1. ***Jump up to:[a] [b] Olkin, I.; Sampson, A. R. (2001-01-01), "Multivariate Analysis: Overview", in Smelser, Neil J.; Baltes, Paul B. (eds.), International Encyclopedia of the Social & Behavioral Sciences, Pergamon, pp. 10240–10247, ISBN 9780080430768, retrieved 2019-09-02***
2. ***Sen, Pranab Kumar; Anderson, T. W.; Arnold, S. F.; Eaton, M. L.; Giri, N. C.; Gnanadesikan, R.; Kendall, M. G.; Kshirsagar, A. M.; et al. (June 1986). "Review: Contemporary Textbooks on Multivariate Statistical Analysis: A Panoramic Appraisal and Critique". Journal of the American Statistical Association. 81 (394): 560–564. doi:10.2307/2289251. ISSN 0162-1459. JSTOR 2289251.(Pages 560–561)***
3. ***Schervish, Mark J. (November 1987). "A Review of Multivariate Analysis". Statistical Science. 2 (4): 396–413. doi:10.1214/ss/1177013111. ISSN 0883-4237. JSTOR 2245530.***
4. ***T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.***
5. ***KV Mardia; JT Kent & JM Bibby (1979). Multivariate Analysis. Academic Press. ISBN 978-0124712522. (M.A. level "likelihood" approach)***
6. ***Feinstein, A. R. (1996) Multivariable Analysis. New Haven, CT: Yale University Press.***
7. *Hair, J. F. Jr. (1995) Multivariate Data Analysis with Readings, 4th ed. Prentice-Hall.*
8. ***Johnson, Richard A.; Wichern, Dean W. (2007). Applied Multivariate Statistical Analysis (Sixth ed.). Prentice Hall. ISBN 978-0-13-187715-3.***
9. ***Schafer, J. L. (1997) Analysis of Incomplete Multivariate Data. CRC Press. (Advanced)***
10. ***Sharma, S. (1996) Applied Multivariate Techniques. Wiley. (Informal, applied)***
11. ***Izenman, Alan J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics. New York: Springer-Verlag. ISBN 9780387781884.***
12. ***"Handbook of Applied Multivariate Statistics and Mathematical Modeling | ScienceDirect". Retrieved 2019-09-03.***